
College Football Statistics Project

Cody Faircloth

Who Am I?

- Name: Cody Faircloth
- Age: 25 years old
- Occupation: Data Engineer
- Education: BS Computer Science, minor in Software Engineering

Why Tech?

- Childhood Interest
- Constraints Choosing a College Major
- The Process and Continuous Learning

Technologies Used

- Programming Language
 - Python
- Relational Database Management System (RDBMS)
 - MySQL
- Containerization Platform
 - Docker
- Docker Images
 - Alpine
 - Apache Airflow
 - Apache Spark
- Python Libraries
 - Airflow
 - Dash
 - DynaConf
 - SQLAlchemy
 - SSHOperator
 - Pandas
 - Pendulum
 - Plotly Express
 - PySpark

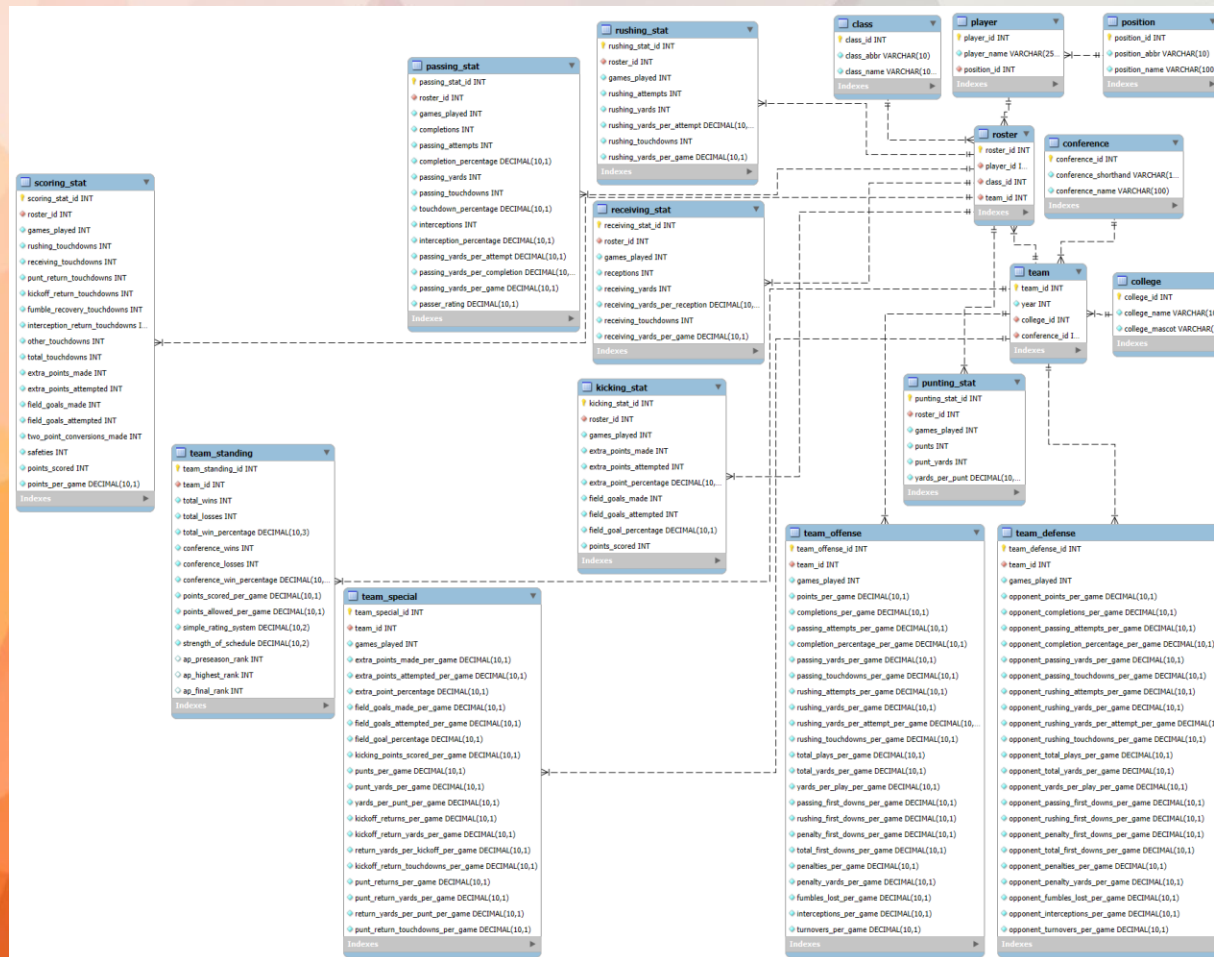
Goals

- Scalable Dashboard Similar to ESPN
- Identify Trends by Team, Conference, and Nationally

Schema Design Process

- Complex Dataset
- Needs Scalability
- Complex Schema
- Practical Normalization

Finalized Schema



ETL

- Scalable
- Separation of Concerns
- Relatively Efficient

```
# Functions to build dataframes for ETL process
> def build_master_college_conf(years): ...

> def build_master_players(year, teams): ...

> def build_conferences(master): ...

> def build_colleges(master): ...

> def build_teams(master, colleges, conferences): ...

> def build_positions(master): ...

> def build_classes(master): ...

> def build_players(master, positions): ...

> def build_rosters(master, players, teams, classes): ...

> def build_rushing(years, rosters): ...

> def build_receiving(years, rosters): ...

> def build_passing(years, rosters): ...

> def build_scoring(years, rosters): ...

> def build_punting(years, rosters): ...

> def build_kicking(years, rosters): ...

> def build_team_offense(years, teams): ...

> def build_team_defense(years, teams): ...

> def build_team_special(years, teams): ...

> def build_standings(years, teams): ...

# List of years to process
years = [2024, 2023, 2022, 2021, 2020]
```


Airflow

- Schedule ETL Process
- Run Schema SQL Script
- Run Spark command

```
import pendulum
from airflow.decorators import dag
from airflow.providers.common.sql.operators.sql import SQLExecuteQueryOperator
from airflow.providers.ssh.operators.ssh import SSHOperator

# Create a DAG to run cfb_statistics_schema.sql and cfb_stats_app.py
@dag(
    schedule_interval=None,
    start_date=pendulum.datetime(2025, 1, 1, tz="UTC"),
    catchup=False,
    tags=["cfb_stats", "ssh", "mysql"],
)
def ssh_mysql_cfb_stats_dag():
    # Define the SSH command to run cfb_stats_app.py
    command = """
    export JAVA_HOME=/opt/java/openjdk;
    /opt/spark/bin/spark-submit --master spark://spark-leader-cfb:7077 /opt/spark/work-dir/
    cfb_stats_app.py
    """

    # Create SSHOperator task to run cfb_stats_app.py
    ssh_cfb_stats = SSHOperator(
        ssh_conn_id="ssh_leader", task_id="ssh_task", command=command, cmd_timeout=600
    )

    # Create SQLExecuteQueryOperator task to run cfb_statistics_schema.sql
    cfb_stats_schema = SQLExecuteQueryOperator(
        task_id="cfb_stats_schema",
        sql="/cfb_statistics_schema.sql",
        conn_id="mysql_sys"
    )

    # Schedule tasks
    cfb_stats_schema >> ssh_cfb_stats

ssh_mysql_cfb_stats_dag()
```

Visualizations

- Roster and Stat Tables
- Team Standings in Conference and Nationally
- Stats over Time Graphs
- Stat Distribution

Visualization Examples

College Football Dashboard

Stats by Team

Stats by Conference


National Stats

Select a Year

2024

Select a College

Georgia



2024 Georgia Bulldogs

Conference: Southeastern Conference (SEC)

Player	POS	Class
Ali Hall	DB	SR
Chris Peal	DB	SO
Collin Gill	DB	SO
Dan Jackson	DB	SR
Daniel Harris	DB	JR
Daylen Everette	DB	SR
Demello Jones	DB	SO
Denton Shamburger	DB	JR
Ellis Robinson IV	DB	SO
Henry Waller	DB	JR
Jacob Fleming	DB	JR
JaCorey Thomas	DB	SR
Jake Pope	DB	JR
Joemel Agiero	DB	JR
Julian Humphrey	DB	JR
Justyn Rhett	DB	SO
KJ Bolden	DB	SO
Kyron Jones	DB	JR
Malaki Starks	DB	SR
Maurice Hayes	DB	SO
Ondre Evans	DB	SO
Parker Jones	DB	JR
Justin Greene	DE	SO
Christen Miller	DL	JR
Clinton Barlow	DL	SO

1

6

Stats by Team: Roster and Header

Team Stats

Individual Stats

Team Stats Over Time

Player Stat Distribution

2024 Georgia Offense (Per Game Stats)

GP	Pts	Cmp	PAtt	Cmp%	PYds	PTD	RAtt	RYds	Y/A	RTD	Plays	TYds	Y/P	PFD	RFD	PenFD	TFD	Pen	PenYds	Fum	Int	TO
14	31.5	24	36.6	65.5	281	2.1	30.6	124.4	4.1	1.7	67.3	405.4	6	12.9	7.3	1.6	21.9	5.7	53.6	0.5	0.9	1.4

2024 Georgia Defense (Opponent Per Game Stats)

GP	Pts	Cmp	PAtt	Cmp%	PYds	PTD	RAtt	RYds	Y/A	RTD	Plays	TYds	Y/P	PFD	RFD	PenFD	TFD	Pen	PenYds	Fum	Int	TO
14	20.6	17.9	30.4	58.9	200.2	1	34.9	129.6	3.7	1	65.4	329.9	5	8.9	7	1.9	17.8	6.4	49.3	0.8	0.6	1.4

2024 Georgia Special Teams (Per Game Stats)

GP	XPM	XPA	XP%	FGM	FGA	FG%	KPts	Punt	PuntYds	Y/Punt	KR	KRYds	Y/KR	KRTD	PR	PRYds	Y/PR	F
14	3.4	3.4	100	1.5	1.6	91.3	7.9	3.3	154.1	46.9	0.9	13.3	15.5	0	1.5	12	8	

Stats by Team: Team Stats Tab

Visualization Examples (cont.)

College Football Dashboard

Stats by Team

Stats by Conference

National Stats

Select a Year

2022

Select a Conference

SEC



2022 Southeastern Conference (SEC)

College	TV	TL	TW	OW	CL	CW	PS/G	PA/G	SRS	SOS	APPre	APHigh	APPost
Georgia	15	0	1	8	0	1	41.1	14.3	25.48	6.28	3	1	1
Alabama	11	2	0.846	6	2	0.75	41.1	18.2	19.66	4.58	1	1	5
Tennessee	11	2	0.846	6	2	0.75	46.1	22.8	17.9	4.98		2	6
LSU	10	4	0.714	6	2	0.75	34.5	22.5	13.29	6.14		6	16
Mississippi State	9	4	0.692	4	4	0.5	31.6	23.1	11.84	5.3		16	20
Ole Miss	8	5	0.615	4	4	0.5	33.5	25.5	8.36	4.05	21	7	
South Carolina	8	5	0.615	4	4	0.5	32.2	28.8	7.71	4.17		20	23
Arkansas	7	6	0.538	3	5	0.375	32.5	30.6	5.95	5.1	19	10	
Kentucky	7	6	0.538	3	5	0.375	28.4	19.2	6.05	4.74	20	7	
Florida	6	7	0.462	3	5	0.375	29.5	28.8	6.92	7		12	
Missouri	6	7	0.462	3	5	0.375	24.8	25.2	4.95	4.49			
Auburn	5	7	0.417	2	6	0.25	24.8	29.5	3.9	7.4			
Texas A&M	5	7	0.417	2	6	0.25	22.8	21.2	3.21	3.13	6	6	
Vanderbilt	5	7	0.417	2	6	0.25	24.6	36	-1.14	5.11			

Stats by Conference:
Standings and Header

Team Stats

Individual Stats

Conference Stats Over Time

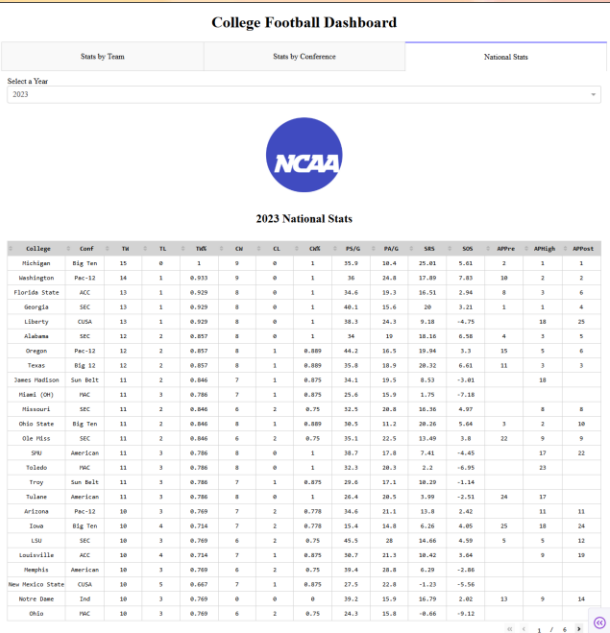
Team Stat Distribution

2022 SEC Rushing

Player	College	GP	Att	Yds	Y/A	TD	Y/G
Quinshon Judkins	Ole Miss	13	274	1567	5.7	16	120.5
Raheim Sanders	Arkansas	13	222	1443	6.5	10	111
Devon Achane	Texas A&M	10	196	1182	5.6	8	118.2
Re'Mahn Davis	Vanderbilt	12	232	1042	4.5	5	86.8
Tank Bigsby	Auburn	12	179	970	5.4	10	80.8
Zach Evans	Ole Miss	12	144	936	6.5	9	78
Jahmyr Gibbs	Alabama	12	151	926	6.1	7	77.2
Christopher Rodriguez Jr.	Kentucky	8	175	904	5.2	6	113
Jayden Daniels	LSU	14	186	885	4.8	11	63.2
Jaylen Wright	Tennessee	13	146	875	6	10	67.3
Montrell Johnson	Florida	13	155	841	5.4	10	64.7
Kenny McIntosh	Georgia	15	150	829	5.5	10	55.3
Daijun Edwards	Georgia	15	140	769	5.5	7	51.3
Cody Schrader	Missouri	13	170	745	4.4	9	57.3
Jabari Small	Tennessee	13	157	734	4.7	13	56.5
Trevor Etienne	Florida	13	118	719	6.1	6	55.3
Robby Ashford	Auburn	12	153	710	4.6	7	59.2
Jarquez Hunter	Auburn	12	104	668	6.4	7	55.7
Jase McClellan	Alabama	13	112	655	5.8	7	50.4
Anthony Richardson	Florida	12	103	654	6.3	9	54.5
K.J. Jefferson	Arkansas	11	158	640	4.1	9	58.2
Jaxson Dart	Ole Miss	13	128	614	4.8	1	47.2
Kendall Milton	Georgia	13	85	592	7	8	45.5
Brady Cook	Missouri	13	139	585	4.2	6	45
Jo'quavious Marks	Mississippi State	12	113	582	5.2	9	48.5

Stats by Conference:
Individual Stats

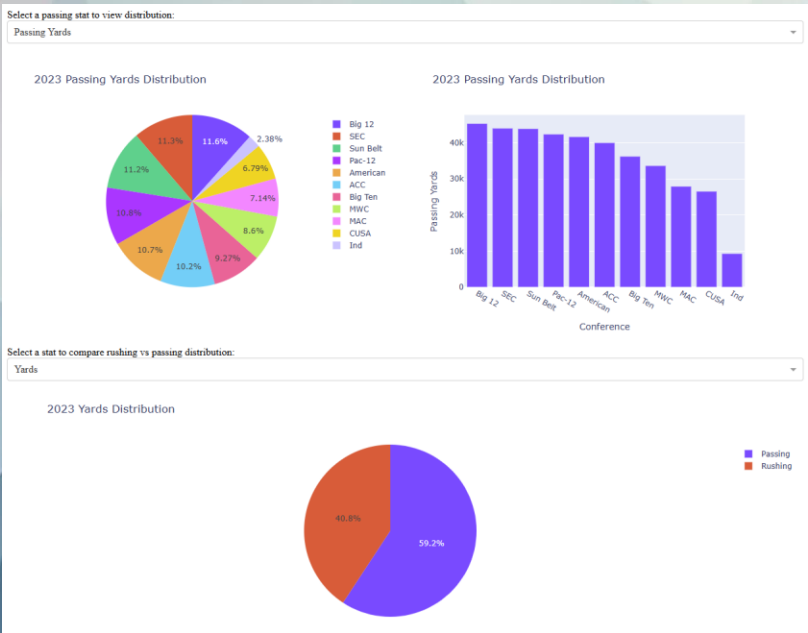
Visualization Examples (cont.)



National Stats: Standings and Header



Stats by Team: Team Stats over Time



National Stats: Conference Stat Distribution

Conclusions

- Dashboard displays accurate data
- Dashboard data retrieval is efficient
- Trends Conclusions vary by team
- Distribution of play types favor passing